DOCUMENT RESUME

ED 204 957 EC 133 528

AUTHOR Gersten, Russell M.
TITLE Direct Instruction Programs in Special Education Settings: A Review of Evaluation Research Findings.
PUB DATE Apr 81
NOTE 41p.; Paper presented at the Annual International Convention of The Council for Exceptional Children (59th, New York, NY, April, 1981, Session Th-58).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Disabilities: Elementary Secondary Education: *Evaluation Methods: Program Evaluation: *Research Design: *Research Methodology: *Research Problems
IDENTIFIERS *Direct Instruction Model

ABSTRACT

The paper reviews the state of the art in evaluation research of direct instruction projects with handicapped students. The first section delineates research designs used (including norm referenced comparisons, nonequivalent comparison groups, and quasi experimental designs): major findings: and problems with each. A distinction is made between evaluation research and documentation activities. Documentation is seen to lack the use of comparison standards. The final section examines the definition of the independent variable and considers directions for future research in formative evaluation and instructional dimensions. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DIRECT INSTRUCTION PROGRAMS IN SPECIAL EDUCATION SETTINGS: A REVIEW OF EVALUATION RESEARCH FINDINGS

by

Russell M. Gersten, Ph.D.

Department of Special Education
University of Oregon

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Russell M. Gersten
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at Annual Conference of
Council for Exceptional Children;
New York, New York
April, 1981

Because the phrase Direct Instruction evolved out of the experience of Engelmann, Becker, Carnine and their colleagues' work in the field of compensatory education, most of the early conceptualizations (Becker, 1977; Becker, Engelmann and Thomas, 1974; Bereiter and Engelmann, 1964; Rosenshine and Berliner, 1978) focused on the educational concerns of low income, minority students in the primary grades. The last five years have witnessed a growing realization that many of the principles and concepts of direct instruction have wide relevance for special education populations (Bellamy, Horner and Inman, 1979; Stephens and Rosenshine, in press; Lloyd, et al., chapter ; Gersten and Maggs, 1980; Maggs and Maggs, 1980; Bateman and Carnine, 1977; Engelmann, 1977; Bateman, 1979). This process was no doubt accelerated by the publication of several direct instruction programs targeted for special students (e.g., Corrective Reading, Reading and Arithmetic series).

Direct Instruction shares many features with the task analytic, behavioral approaches commonly utilized in special education, viz. the use of task analysis, a belief in the utility of structured curricula materials, a concern with reinforcement, modelling and shaping of correct responses, and periodic assessment of student performance. Yet there are certain features that distinguish direct instruction from other behavioral approaches. These include, (1) the teaching of "general" case problem solving strategies; (2) a maximal use of oral instruction as opposed to written worksheets; (3) the emphasis on small group as opposed to individualized instruction; (4) a systematic technology of correction procedures to transform student errors into constructive learning experiences.

Programs falling under the rubric "direct instruction" always put emphasis on the curricula materials as well as the teaching procedures. See Lloyd, et al. (in press); Gersten and Maggs (1980); Fulco (1980) for further discussion of these issues.

The present paper began as an attempt to pull together and integrate the various experimental and quasi-experimental studies that have been conducted with special education populations over the last decade in the United States. But, like many other ventures in life, the direction shifted midstream. Evaluation research is a relatively new venture in education (Cooley and Lohnes; VonBach, 1980), and particularly in special education (C.F. Hallahan and Kaufman, 1976), while the focus on individualization of instruction and wariness of standardized tests often create strong resistance to typical evaluation research design (Kennedy, 1980). As I confronted the research studies and documentation activities of many direct instruction projects, it seemed as important to highlight and pinpoint roadblocks conceptual problems and technical difficulties encountered by the authors as to share their findings. So this essay is a little of both - ending with some ideas for small scale evaluation or research projects that can and should be undertaken in the near future.

The first section of this paper focuses on the how aspects of evaluation research - delineating the research designs used in the past, the major findings, and problems with each. Measurement problems encountered by the authors are discussed, along with viable solutions. The major findings of key studies are shared. Finally, the distinction

is made between evaluation research and documentation activities - and the utility of each approach is discussed.

The closing section confronts an issue that often hides in the closet - the what of evaluation research, i.e., the definition of the independent variable. In future research on direct instruction, more clarity and precision is needed in the definition of the term as it is used in each study. Some suggestions from research on implementation of curriculum (Fullan, 1980; Hall and Loucks, 1977; Gersten, Carnine and Williams, in press; Zoref, 1981), instructional dimensions studies (Cooley and Leinhardt, 1980), and formative evaluation (Gersten and Carnine, 1979) are presented as promising directions. Findings on some early studies examine how some direct instruction procedures may need to be adapted to meet the needs of handicapped learners are presented.

EVALUATION RESEARCH

"True" Experiments (Experiments with Random Assignment):

For a variety of reasons, the strongest evaluation research design is one in which relatively large groups of students are randomly assigned to one of two educational treatments, and, after treatment, the mean performance of the two samples on one or more sample measures is analyzed to determine whether one performs at a significantly different level from the other.

Usually, one of the samples is designated as the experimental sample and is taught with the instructional program of interest (in this case Direct Instruction), and the other sample receives more traditional

instruction and is labelled a comparison group. (For detailed rationales see Campbell and Bourch, 1975). For this type of study to offer any valuable information, the tests used must be valid, reliable and sensitive to the goals of the instructional programs (Fullan, 1980; House, Glass, McLean and Walker, 1978). Also, it is extremely important that the evaluation monitor to what extent the new program (in this case Direct Instruction) is really being implemented in the experimental classes (Gall, 1977; Becker and Engelmann, 1978; Fullan, 1980); otherwise, the study may be assessing the impact of a non-event (Charters and Jones, 1973). There should also be some monitoring of the comparison classrooms (Gersten, Becker, and Carnine, 1979) to insure that these classes are not using the experimental teaching method.

If all these conditions are met, and the sample size is large enough (usually at least 15 children per condition), then one has an extremely strong evaluation design. One can be reasonably sure that differences between the samples on the post-tests are due to the educational program and not to demographic variables, placement procedures in a given school district, or other extraneous factors. There are so many factors mitigating against this type of design - refusal of parents and school personnel to randomly assign students to instructional programs, attention of students, difficulties in finding appropriate tests for handicapped learners - that it is a wonder any have been done at all with Direct Instruction and handicapped learners. In fact, two quite respectable studies have been conducted and they will be discussed in some detail.

The first "true experiment" conducted in this field was an investigation of three instructional procedures for "high risk" first graders in a suburban community. Serwer, Shapiro and Shapiro (1973) deemed 62 of the 300 kindergartners in two schools as "high risk" on the basis of "observed difference between potential and achievement with due regard in both those areas to the verbal, performance and visual motor dimensions." (p. 242). Screening was based on teacher questionnaires, two reading tests and two I.Q. tests. All were administered at the end of kindergarten. Note that over 20 percent of the suburban students were targeted as "high risk", though typical estimates of students needing some special education services are in the range of 10 to 12 percent.

Students were randomly assigned to one of four remediation groups - direct instruction (30 minutes a day of Distar Reading), indirect instruction (use of perceptual motor activities for 30 minutes a day, based on the work of Kephart (1964)), both treatments combined, or control. The combined group received 15 minutes of each treatment, the control group no special treatment. In all cases students received the remedial treatment in addition to their regular first grade reading program. The nature of these programs is not specified. There is another anomaly in this study. All students were tested midyear on a variety of auditory-visual, fine motor, and gross motor tests, and were then given "prescriptions". For the remainder of the year, "the Direct group combined with the Distar program and was trained in the deficit modalities as diagnosed by using letter forms and sounds and word forms and sounds." These procedures are clearly not part of direct instruction (see Lloyd, et. al, in press; Bateman, 1975).

Results of the mid year tests indicated a significant difference in the Word Recognition Test (Serwer, 1971); $F_{(3,53)} = 7.98$, $p < .001$. Oddly, post-hoc tests were performed, but it appears that students in the Indirect (Kepart) treatment performed significantly higher than comparison students. No other significant differences were found.

The final post-tests showed a significant difference on the Mathematics subtest of the Metropolitan Achievement Test; $F_{(3,53)} = 3.40$; $p < .05$, seeming to favor the Kephart group over the Direct group. This is a curious finding since Arithmetic was not a target of any of the three treatments. A significant difference was found in the Wrong Endings subtest of the Gates-McKillop favoring the Direct group. The combined and Indirect groups performed significantly better than the Direct or Control students on three of the ten tests of motor skills - Locomotor balance, hand-eye coordination, and writing.

It is unclear what to make of this study. Despite the myriad of tests given, few significant results are found. The huge standard deviations on many of the measures would seem to indicate that the tests rarely assess the goals of the programs and/or that the population is very heterogenous. The only significant findings are: (1) students in the Kephart program perform more highly on two non-accedemic tests (Locomotor Balance, Eye-hand Coordination) and two academically oriented tests (Handwriting, Math) that are not an emphasis of any program. Students in the Direct group perform significantly different on a test of "Wrong Endings." The author's failure to perform post-hoc tests further obscures the results. This terribly flawed study exemplifies the

importance of:

a) offering a clear - and reasonable definition of the target population;

b) carefully defining and implementing the educational treatments (in this case, the direct method does not adhere to many principles of direct instruction during the latter part of the year). Also, most students received an additional hour per day of (unspecified) reading instruction in their regular classrooms;

c) selection of tests to meet the objectives of the programs.

Although the study shows that to some limited extent the Kephart program is effective in helping children with two of the motor skills, little can be concluded concerning how to teach reading to first graders with learning disabilities.

Language Instruction for the Moderately Retarded:

Maggs and Morath (1976) investigated the effects of Distar Language I on moderately to severely retarded children in a state institution. Twenty-eight students, ranging in ages between 6 and 14 with Stanford-Binet I.Q.s ranging between 20 and 45, were randomly assigned to either an experimental group (one hour per day of Distar Language I coupled with behavioral teaching procedures) or a comparison group (one hour per day with the commonly used Peabody Language Kit and supplemented teacher-generated language programs). The program lasted two full years. The experimental group scored significantly higher on the Stanford Binet; the mean gain for this group was 22 1/2 "mental age months" for the 24 months in the program, approximating normal intellectual growth rates.

The control group, however, gained only 7 1/2 mental age months in the 24 months of the program, showing the continual intellectual deterioration usually observed in institutional settings, and pointing up graphically the ever-widening gap between "retarded" and "normal".

Children were also tested on a series of cognitive tasks based on the work of Piaget and Bruner to insure that children were learning to generalize the concepts they had learned in the language programs. Again, significant differences were found favoring the experimental group. The authors wished to explore whether concepts learned purely by verbal instruction would transfer to a situation in which the concepts had to be demonstrated by manipulating concrete objects. The results showed a significantly better post-test performance by the experimental subjects, indicating their skills went beyond mere rote learning and that concepts taught verbally, with only two-dimensional visual material could be applied readily to three-dimensional concrete objects.

This field study documented that much of the intelligent behavior assessed in tests like the Stanford Binet can be taught through extensive systematic instruction, even to moderate and severely retarded children. Children in the Maggs and Morath (1976) study for two years approached the "normal" level of growth. This point will be of interest when quasi-experimental designs are discussed.

Teaching Reading to Learning Disabled Students:

Lloyd, et. al (1980, in press) randomly assigned 23 learning disabled students in the intermediate grades to one of three different classrooms. The two experimental classes utilized direct instruction progra

in reading -- decoding basics (Engelmann, et. al), a proficiency version of the Corrective Reading Program, Levels A and B, for word attack skills and thinking basics for reading comprehension. No direct instruction arithmetic program was used. An applied behavior analysis approach to monitoring student progress, involving specific probe sheets for academic skill areas, was also used in both reading and arithmetic. (This system is not part of conventional direct instruction programs).

Students in the comparison classroom received individual and small group instruction in various academic areas (reading, language arts, math, spelling) as well as training in psychological processes believed to be related to learning. The program was considered typical of traditional special education programs for the learning disabled. Students in both experimental and comparison classes were in a full day program.

Students were assessed on (a) the Wide Range Achievement Test (Jastak and Jastak, 1976), a measure of Reading (decoding), Arithmetic (computation), and Spelling; (b) the Gilmore Oral Reading Test (Gilmore and Gilmore, 1968), and (c) Slosson Intelligence Test (Slosson, 1971). The Gilmore includes measures of Reading Accuracy (Decoding), Reading Rate, and Comprehension. The Slosson was used to assess general language competence.

Results indicated statistically significant differences favoring the experimental group in measures of decoding and comprehension (see Table 2).

The difference in reading rate was on the borderline of statistical significance ($t = 1.68$, $p = .054$) and is clearly educationally significant (magnitude of effect of .66 posted 55 units). On the other hand,

Table 1.

COMPARISON OF L.D. STUDENTS WITH ONE YEAR OF CORRECTIVE READING PROGRAMS IN DECODING AND COMPREHENSION WITH STUDENTS IN TRADITIONAL PROGRAMS

Outcome Measure	Groups								
	E_1		E_2		Comparison				
	M	SD	M	SD	M	SD	t	p[a]	relative advantage
DECODING									
WRAT Reading S.S.	83.29	10.69	81.75	9.67	73.88	8.77	1.897	.036	.82
Gilmore Accuracy Index	72.15	21.05	53.67	16.40	47.82	19.82	1.682	.054	.64
Gilmore Rate/Minute	60.86	30.33	64.50	29.51	44.00	12.77	1.572	.066	.70
Gilmore Comprehension Index	87.13	27.22	65.04	26.48	54.41	24.70	1.769	.046	.71
Slosson Intelligence Text	85.86	8.17	85.63	6.38	80.13	6.01	1.74	.048	.75
ARITHMETIC SKILLS									
WRAT Arithmetic S.S.	83.86	11.14	82.13	2.20	80.88	4.59	0.658	.259	.30

(Adapted from Lloyd, et al, (1980; in press)

[a]Directional test

[b]Inpooled SD units

the difference in WRAT arithmetic is clearly non-significant; t = .658, p = .26. This offers further confirmation of the effectiveness of the direct instruction programs (Campbell and Fiske, 1959), since no direct instruction math program was implemented.

This study documents the relative superiority of direct instruction reading programs in the domains of decoding and comprehension. It is one of the few experimental studies to use learning disabled students, and one of the few studies to document the effectiveness of the newly developed comprehension programs. On the technical side, this is one of the best designed special education research studies - with its use of random assignment of children to treatment, the broad range of measures used, its relatively clear definition of the handicapped, and its assessment of performance in both target (reading) and non-target (arithmetic) academic areas. Lloyd, et. al's (in press) use of improvised standard scores on the Gilmore is a bit more questionable, since there is considerable controversy concerning the use of grade equivalent scores (Horst, Tallmadge and Wood, 1975) and mental age - to - chronological age ratios (Matarzzo, 1972). There is no reason to assume the procedure systematically biases the results.

Two issues raised by the authors deserve further elaboration and discussion. Though the experimental group consistently surpassed the comparison group, one year of direct instruction reading was <u>not</u> sufficient to bring those students to a "normal" level of performance (That would represent mean standard scores of 100 on the measures). In fact, there is evidence in a small drop in measured I.Q. for the two experi-

mental classrooms - from 90.4 to 87.1 for one group and 87.5 to 85.6 for the other. But this drop is appreciably less than the drop for the comparison group. The drop may be artifactual - students were pre-tested on the WICS, which contains performance and language items, and post-tested on the Slosson, which is primarily verbal, and hence may underestimate scores of learning deficient students. As the authors state, "these data do not argue that . . . the procedures "cured" learning disabilities," (in ten months), "but they do indicate a direction that clearly merits further direction." (p. 31).

Another direction that needs to be pursued is a more fine-grained analysis of exactly what the instructional program was like. It is described as "a hybrid behavioral model" using both behavioral principles of programming and direct instruction curricula and teaching techniques. Future field studies could explore the relative importance of these components - how important are probes and daily monitoring procedures? What adaptations did the teachers make from specified formats - and did these aid or hinder student performance?

Quasi-Experimental Designs:

The last two studies discussed above - which involved random assignment of subjects to educational treatment - though imperfect, offer some degree of methodological rigor to the questions they pose. Unfortunately, it is often impossible or infeasible to <u>randomly</u> assign children to educational programs. There are a host of ethical and logistical reasons why most school systems will not allow selection of educational programs for handicapped students to be determined by the toss of a coin or a flip

through a random numbers table. Even Lloyd and Maggs, principal authors of the two previous studies, have been unable to achieve random assignment in subsequent research (Lloyd, personal communication; Maggs, personal communication).

Two alternatives exist - single subject research (such as the example discussed by Anderson-Inman in this presentation) and quasi-experimental designs. The latter will be the central focus in this section.

Three major types of quasi-experimental designs appear to be appropriate for direct instruction (Cook and Campbell, 1979), evaluation research with special education populations:

1. Non-equivalent comparison groups

2. Norm-referenced comparisons (Becker and Engelmann, 1976; Tallmadge, 1977).

Each has its limitations; yet, used with care, these can be useful designs, and the first two have been used to provide some valuable information on the effects of direct instruction programs on special education students.

Non-Equivalent Comparison Group:

A recent study by L' E. Stein and Goldman (1980) contrasted the effectiveness of two phonics-based programs - Distar and Palo Alto - on 63 "primary grade children with reading problems." between the ages of 6 and 8. The children possessed average I.Q.s ($\bar{X}$ = 100.1). All the children were formally diagnosed as possessing minimal brain dysfunction, but the authors (rightly in our opinion) felt that "the term was not particularly relevant to each specific child's curriculum, which was based on

pre-test needs as measured through academic assessment" (p. 53).

L' E. Stein and Goldman do a careful job of articulating their two treatment programs, and their rationale for selection and both are phonics-based, since over 15 years of research has shown the superiority of phonics-based instruction for beginning readers (Chall, 1979; Wahlberg, 1980). The major differences between Distar and Palo Alto, according to the authors, are:

1. Distar's insistence on complete mastery of each component skill in the program before continuing on, while Palo Alto calls for less complete mastery.

2. Distar teaches decoding strategies, whereas Palo Alto relies on both systematic decoding and contextual cues (Gill, 1974).

3. Palo Alto allows for greater teacher latitude in directing child behavior than the scripted Distar lessons.

4. On a more theoretical level, Palo Alto is heavily rooted in cognitive theories that look at the teacher as a facilitator of active learning (Beard, 1969; Gibson and Levin, 1975); in Distar, the teacher directs the early stages of the learning process (Engelmann and Carnine; in press; Stephens and Rogersboro). No details are given on assignment of students to program. Thus - as with all non-equivalent group designs - there is always a fear that the two samples were not equivalent at the beginning of treatment, and the results somewhat biased. However, the researchers scrupulously attempted to insure that the groups were comparable and no significant differences were found on entry scores on either the reading recognition or reading comprehension sub-tests of the

Peabody Individual Achievement Test (PIAT) or measured I.Q.s.

Students were post-tested on both subtests of the PIAT and a significant difference was found; $t = 5.27$, df - 56.5; p .001.). This difference favored the Distar group. The mean gain for the Distar group corresponded to 15 months for 9 months of instruction; the mean gain for the Palo Alto students was approximately 7 months. The differences were present in both reading recognition and comprehension measures of the PIAT. (Though it should be mentioned that primarily literal as opposed to inferential comprehension skills are tested at the primary level.) The authors reanalyzed the results using a rather roundabout inefficient version of analysis of covariance and found similar results.

On the positive side, the study looks at two commonly used special education programs and articulates the three salient variables of interest - complete versus incomplete mastery learning, tightly scripted procedures and phonics strategy only versus phonics plus contextual approaches towards teaching decoding. It is unfortunate that no classroom observations were done during the teaching sessions to assess a) the extent to which teachers were following the procedures specified by the curriculum programs and b) quantitative differences in the aforementioned variables. Another crucial variable would be amount of practice on each skill (including review and delayed practice exercises). Procedures for this type of research are discussed in Meyer (1979) Gall (1977) and Gersten, Carnine and Williams (in press).

The L'E. Stein and Goldman (1980) study is the only study using direct instruction to utilize the non-equivalent comparison group design. In this case, the authors did much to insure that the two samples were

comparable - though there are many other studies in both regular education and special education where the two samples were not comparable at the beginning and thus no conclusions at all can be drawn (see Stebbins, et. al, 1977, Volume IV A for the classic discussion; also Campbell and Boruch (1977) and House, et. al, 1978).

Norm-Referenced Comparisons:

This is a relatively new design originating in the work of Horst, Tallmadge and Wood (1975) in Title I evaluations. In many ways, this is an easy design to implement. Students are pre-tested and post-tested on a well-normed, reliable standardized achievement test, whose content matches the objectives of their curriculum program. The gain against the norm group is assessed for statistical and educational significance. Technically, the standardization sample of the test should be comparable to the target population - so Title I students should be compared to other low income students, special education students to other handicapped students, etc. (Tallmadge, 1977). In reality, virtually every standardized test in use is normed on a primarily non-handicapped population, so evaluators have no choice but to use these. The situation is slowly changing; the State of Oregon is in the process of collecting normative data on a curriculum-referenced instrument for Trainable Mentally Retarded students; a consortium of early childhood researchers has developed an assessment instrument, the API, for use only with handicapped infants and preschoolers.

But in most situations one has only the non-handicapped standardization samples to use as comparison standards, and it is often too stringent

a standard to use. For example, in both the Maggs and Morath (1976) and Lloyd, et. al (1980) studies, there are slight, non-significant drops against the norm group.

To be honest, norm-referenced comparisons are often misused in special education. But there are two examples I wish to discuss, both of which this author has been directly involved.

Longitudinal Development of Moderately Retarded Children in a Direct Instruction Program:

Gersten and Maggs (1981) used a norm-referenced evaluation design to assess the cognitive and academic progress over 5 years of a sample of twelve adolescents and preadolescents whose initial I.Q.s would place them in the high moderate range of retardation. The students were engaged in the Distar Language programs for the entire 5 years; after 18 months of language instruction, when they had mastered the basic language competencies and ability to follow instructions in an academic setting, they began Distar Reading I. Each lesson (reading or language) lasted 30 minutes per day and was supplemented by 30 minutes a day of independent seatwork. At the end of the five years the students had completed most of the lessons in the Language III and Reading III programs. The Level III programs contain material intended for average third graders.

Since no comparison group was feasible, the standardization sample of the Stanford-Binet was used as a comparison standard for the evaluation, students were tested on the Binet upon entry, and again 4 years 9 months later, and the mean growth in standard score units (i.e., I.Q. points) was assessed for statistical and educational significance (Tallmadge, 1977), after a correction for regression was made (Campbell and

Erlebacher, 1970). The mean I.Q. at entry was 41.9 (SD = 2.6). After 5 years of the program the mean I.Q. was 50.6 (SD = 5.4). If there was only a regression effect operating, the mean I.Q. at post-test would have been 44.8; there was a gain of .36 standard deviation units beyond what would be expected from regression. This gain was both statistically significant, and educationally significant. The moderately retarded children in this sample were gaining at significantly faster rate than their non-handicapped peers even though typically these children would be expected to lose against the norm group. These students were also tested on Australian standardized tests in Reading and Language; the mean scores were comparable to an end-of-third grade level in language and early third grade level in reading. The children appeared to be close to mastering the basic functional literacy skills essential for vocational training and independent living.

Ideally, the progress of these students would have been compared to a norm group of moderately handicapped students; but such data is currently unavailable. The standardization sample of a well named test such as the Stanford Binet or the Comprehensive Test of Basic Skills is a legitimate if overly stringent comparison standard, and one that is often utilized in special education evaluation. It makes special sense for samples of mildly handicapped students where "normalization" in academic areas is an immediate goal.

The Low I.Q. Study:

Perhaps the most fascinating data comes from secondary analyses of achievement data from the Follow Through Project, a national project for

income students in the primary grades. Students come from twelve communities, including inner city areas in New York and Washington, D.C., the Rosebud Sioux reservation in South Dakota, and rural communities such as Flippin, Arkansas. The general effectiveness of this program in academic areas has already been documented (Becker, 1978; Bereiter and Kurland, in press; Stebbins, et. al, 1977). The purpose of these secondary analyses was to look at the effect of direct instruction and mastery learning on students entering with low cognitive skills, i.e., low I.Q.s. Gersten, Heiry, Becker and White (1981) looked at the yearly achievement test profiles for low income children in the Follow Through programs who entered with I.Q.s on the Slosson Intelligence Test of 70 or below. These children would normally qualify for special education services, but because of the model's belief that students should be taught in regular classrooms, the children were de facto "mainstreamed." Between 1968 and 1977, over 4,000 students received four years of intensive small group instruction in Distar Reading, Language and Arithmetic in their regular classrooms; Inservice training for the teachers and paraprofessionals emphasized that teachers were responsible for the growth of all children, including the lower performers and that with adequate practice, feedback, and reinforcement, the low performers could learn at a normal rate. I.Q. scores were never used for placement purposes or referal for special services; placement and grouping were determined solely by criterion-references tests. The Slosson was used to measure growth in general language competence (Becker, 1977). The Metropolitan Achievement Test and Wide Range Achievement Test were used to assess progress in reading and mathematics.

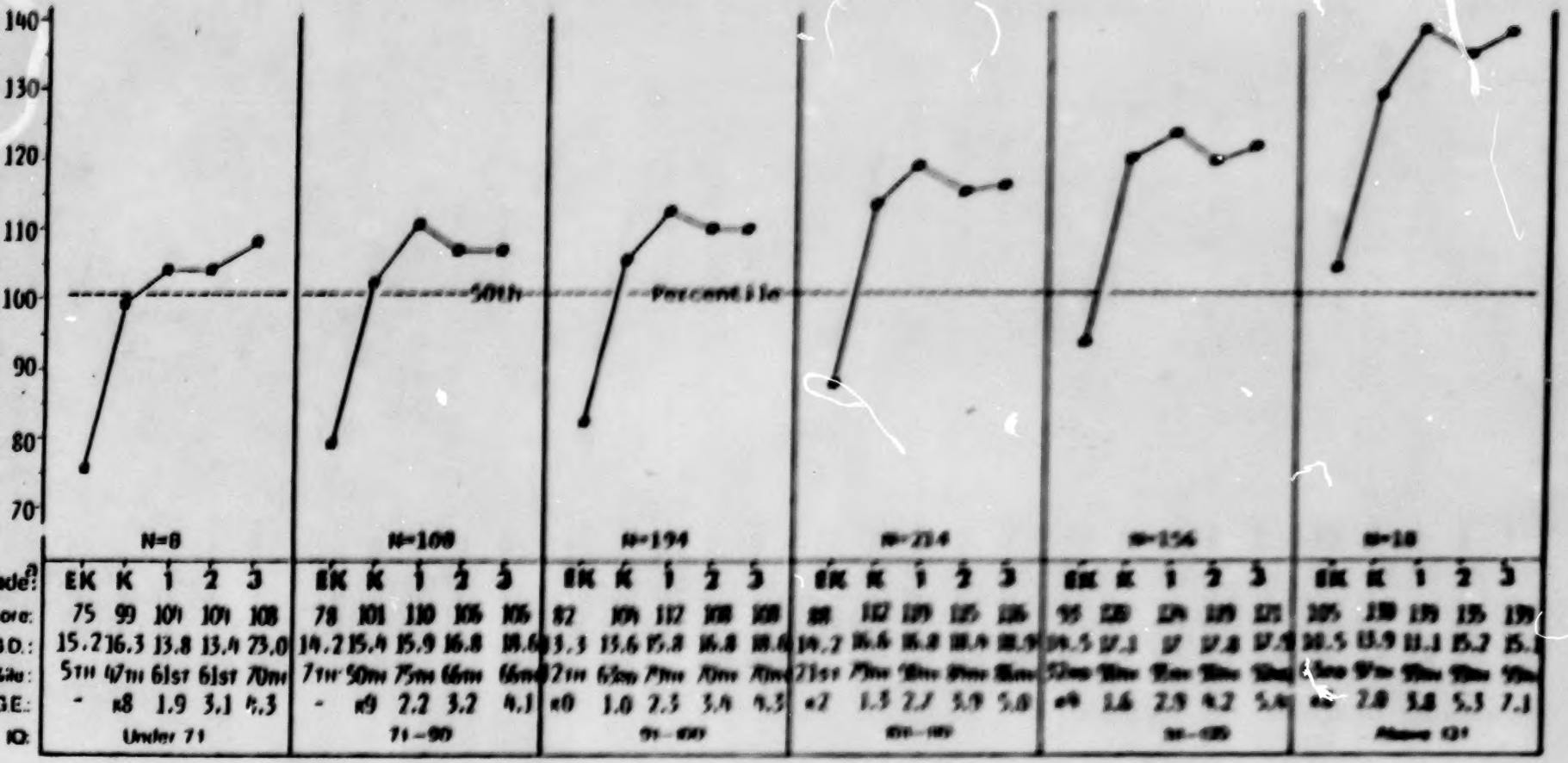
In analyzing the longitudinal achievement data, the hypothesis was that the yearly academic growth rate would be no different for these children than the others in the program. Though the low I.Q. students enter kindergarten with lower academic skill levels, and leave the third grade with lower skill levels than their normal or high I.Q. peers, we believed that their yearly growth - their ability to profit from schooling - would be roughly the same as their peers. In statistical terms, we hypothesized no significant interaction in the 6 x 4 (I.Q. block x time of test) mixed analysis of variance.

Results showed that for children in the full program from kindergarten to third grade, no significant interactions were found on WRAT Reading (a measure of decoding or word identification) or Total Math on the Metropolitan Achievement Test (a measure of math problem solving, concepts and computation). Considering the large sample sizes of 594 and 1096 respectively, and high power, there is reasonably clear evidence of a non-significant effect. Figures 1 and 2 demonstrate yearly gains for the low I.Q. block (70 and below), those with I.Q.s between 71 and 90, and the entire sample on WRAT Reading and MAT Math. On WRAT Reading, the low I.Q. students are virtually at norm levels (the 47th percentile) by the end of kindergarten and continue to make slow but steady growth. By the end of third grade performance corresponds to the 70th percentile, or a 4.3 grade level. A similar pattern is found for the larger 70-90 I.Q. group. For MAT Math, a comprehensive test of problem solving, concepts and conputation, the growth rate for both the below 70 and 70-90 I.Q. blocks corresponds to 1.0 Grade Equivalent units

Figure 1

WRAT Reading

Longitudinal Progress by IQ Block for Children in Entering Kindergarten (EK) Class (N=492)



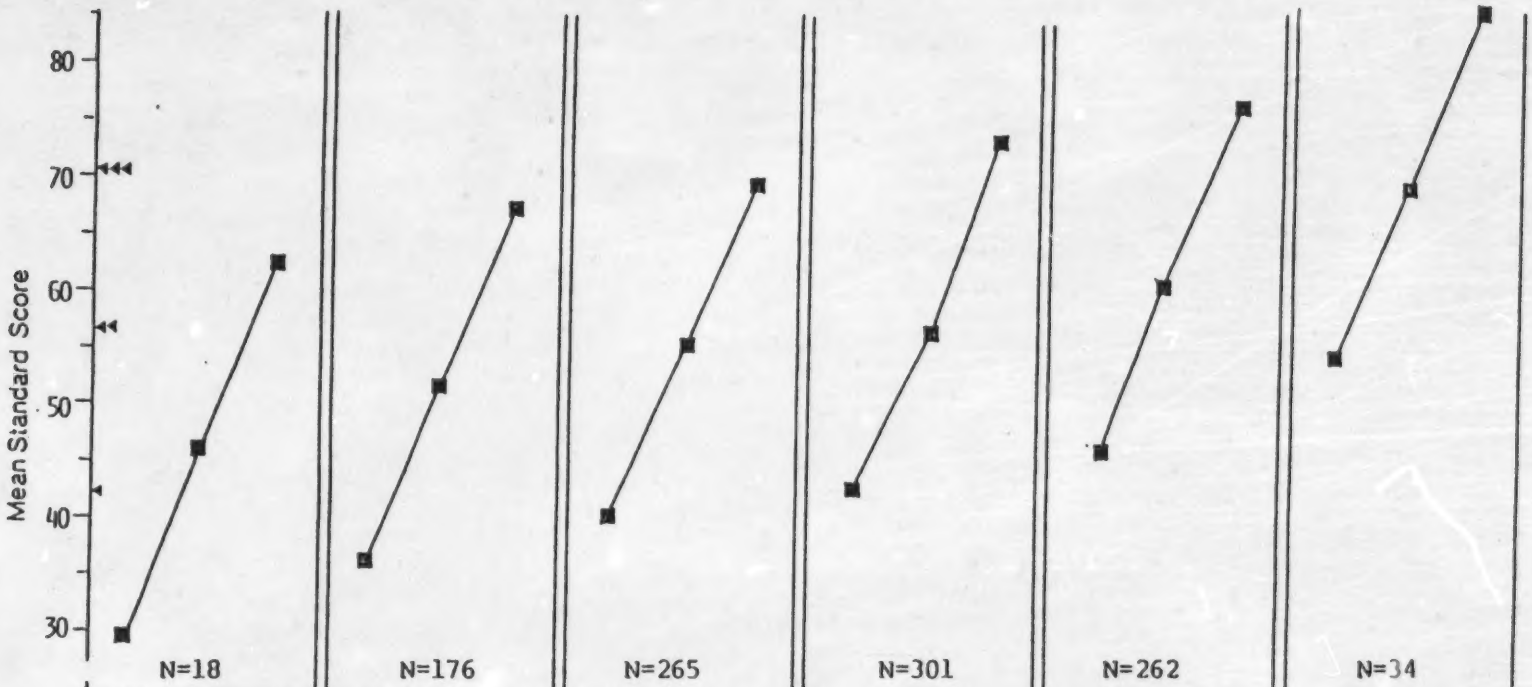
Grade[a]	EK	K	1	2	3	EK	K	1	2	3	EK	K	1	2	3	EK	K	1	2	3	EK	K	1	2	3
Score:	75	99	101	101	108	78	101	110	105	105	82	109	112	108	108	88	112	119	115	116	95	120	124	119	121	105	130	139	135	139
S.D.:	15.2	16.3	13.8	13.4	13.0	14.2	15.4	15.9	16.8	18.6	13.3	15.6	15.8	16.8	18.6	14.2	15.6	15.8	16.9	18.9	14.5	17.1	17	17.8	17.5	10.5	13.9	11.1	15.2	15.1
Rank:	5TH	47TH	61ST	61ST	70TH	7TH	50TH	75TH	66TH	66TH	2TH	63RD	79TH	70TH	70TH	21ST	79TH	86TH	84TH	84TH	52ND	91ST	91ST	91ST	91ST	95TH	99TH	99TH	99TH	99TH
SE:	-	.8	1.9	3.1	4.3	-	.9	2.2	3.2	4.1	.0	1.0	2.3	3.4	4.3	.2	1.3	2.7	3.9	5.0	.9	1.6	2.9	4.2	5.4	.6	2.0	3.8	5.3	7.1
IQ:	Under 71					71–90					91–109					110–119					120–129					Above 129				

[a]All testing performed at end of academic year, except for EK

Figure 2

MAT Total Mathematics

Longitudinal Progress by IQ Block during 2nd and 3rd Grade for Children in Entering-Kindergarten (EK) Sites (N=1,056)



Grade[a]	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Std.Score:	29.2	46.1	62.6	35.8	51.4	67.7	40.0	55.2	69.7	42.6	56.9	73.5	45.6	60.2	76.4	54.4	69.2	85.1
S.D.:	7.3	10.1	9.5	8.0	8.5	10.6	9.0	8.2	9.7	9.2	8.8	10.7	10.8	10.0	11.6	10.6	8.5	8.1
%ile:	10TH	10TH	24TH	30TH	22ND	39TH	46TH	38TH	46TH	51ST	59TH	60TH	57TH	62ND	68TH	82ND	84TH	88TH
G.E.:	1.2	2.0	3.2	1.5	2.3	3.5	1.7	2.5	3.7	2.0	2.3	4.0	2.0	3.1	4.3	2.4	5.2	5.7
IQ:	Under 70			70–90			91–100			101–110			111–130			Above 131		

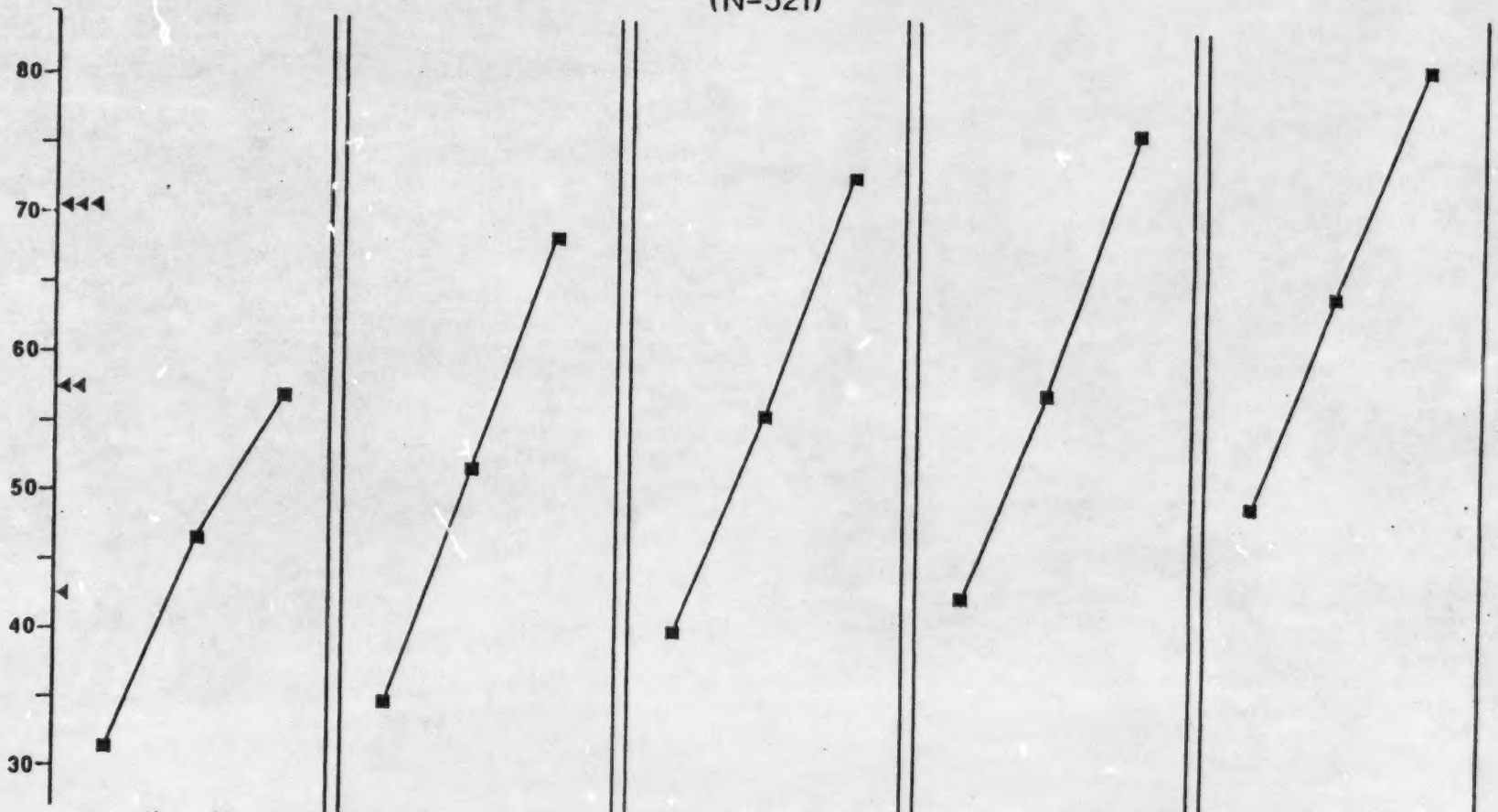
[a]All testing was done at the end of the academic year

◂ Grade 1 National Median
◂◂ Grade 2 National Median
◂◂◂ Grade 3 National Median

MAT Total Mathematics

Longitudinal Progress by IQ Block during 2nd and 3rd Grade for Children in Entering-First(EF) Sites [b]

(N=521)



	N = 20			N = 203			N = 175			N = 83			N = 40		
Grade:[a]	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Std.Score:	31.3	46.4	57.3	34.6	51.5	68.3	39.5	55.3	72.5	42.1	57.3	76.3	48.8	63.8	80.6
S.D.:	7.5	7.5	9.5	7.7	6.9	8.7	8.8	7.7	9.4	9.4	7.4	10.8	14.0	10.3	10.5
%ile:	14th	10th	12th	25th	22nd	40th	41st	38th	57th	50th	50th	68th	68th	77th	78th
G.E.:	1.2	2.0	2.7	1.5	2.3	3.6	1.7	2.5	3.9	1.8	2.7	4.3	2.2	3.4	4.8
IQ:	Under 71			70–90			91–100			101–110			111–130		

[a] All testing done at end of academic year

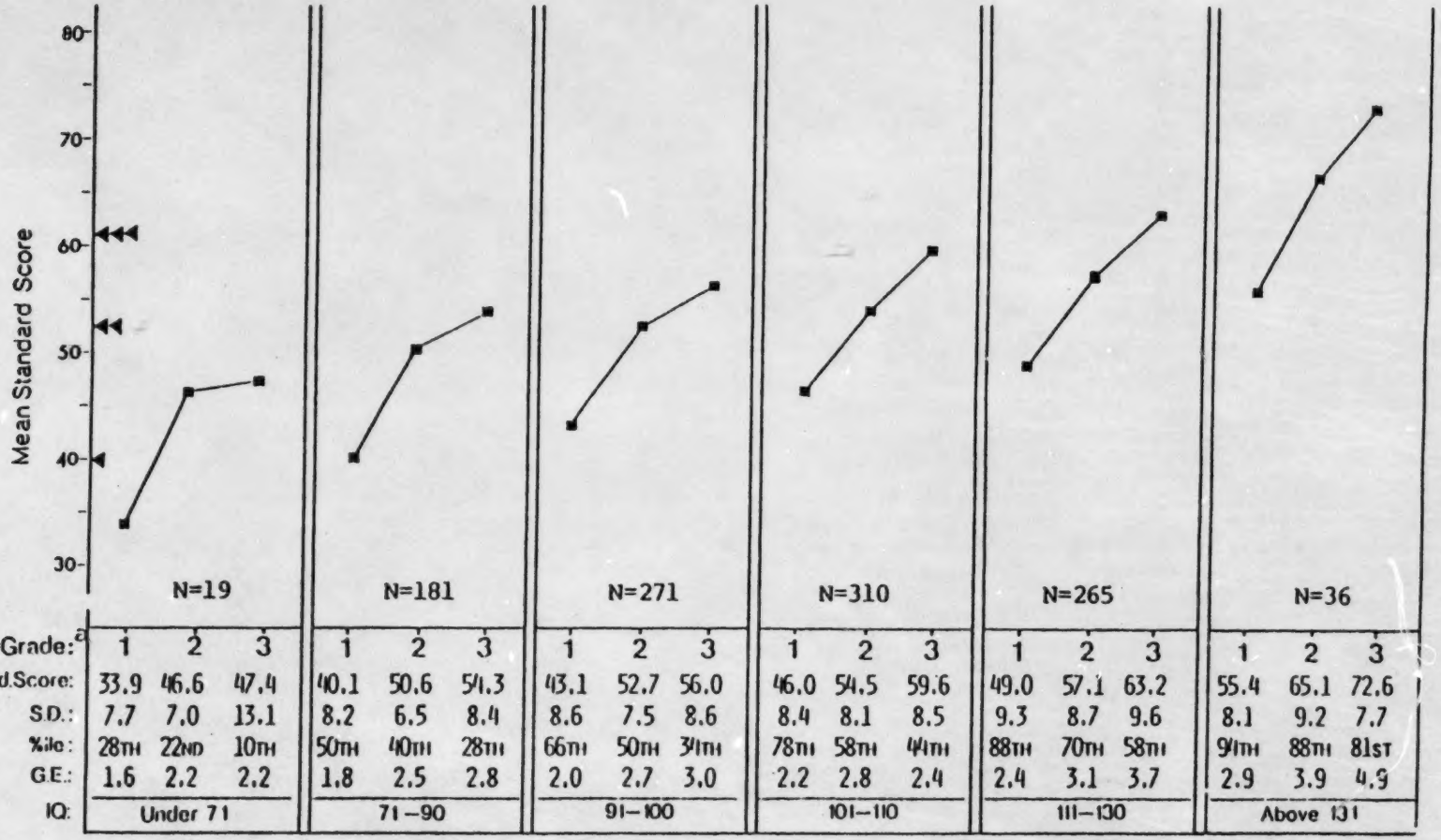
[b] During Grade 3 the lowest and highest IQ blocks gained significantly less than the other IQ blocks

◂ Grade 1 National Median
◂◂ Grade 2 National Median
◂◂◂ Grade 3 National Median

Figure 4

MAT Total Reading

Longitudinal Progress by IQ Block during 2nd and 3rd Grade for Children in Entering-Kindergarten (EK) Sites[b] (N=1,082)



Grade:[a]	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Std.Score:	33.9	46.6	47.4	40.1	50.6	51.3	43.1	52.7	56.0	46.0	54.5	59.6	49.0	57.1	63.2	55.4	65.1	72.6
S.D.:	7.7	7.0	13.1	8.2	6.5	8.4	8.6	7.5	8.6	8.4	8.1	8.5	9.3	8.7	9.6	8.1	9.2	7.7
%ile:	28TH	22ND	10TH	50TH	40TH	28TH	66TH	50TH	31TH	78TH	58TH	41TH	88TH	70TH	58TH	91TH	88TH	81ST
G.E.:	1.6	2.2	2.2	1.8	2.5	2.8	2.0	2.7	3.0	2.2	2.8	2.4	2.4	3.1	3.7	2.9	3.9	4.9
IQ:	Under 71			71–90			91–100			101–110			111–130			Above 131		

[a] All testing was performed at end of academic year

[b] During Grade 3, the highest IQ block gained significantly more than the lowest block (which reflects the verbal content of the SIT and the Total Reading subtest of the Elementary Level MAT)

◂ Grade 1 National Median
◂◂ Grade 2 National Median
◂◂◂ Grade 3 National Median

for each year in school (see Figure 2).[1] The only significant interaction was found on the MAT Total Reading test during the third grade (Elementary). MAT Total Reading is primarily a measure of reading comprehension and vocabulary; here, the low I.Q. group sample gained significantly less than the higher I.Q. children (Figure 4).

The average gain on the Slosson Intelligence Test, a shortened version of the Stanford Binet, was 8 I.Q. points, maintained over the entire four years, for the entire sample of 717 children. The two lowest I.Q. blocks gained significantly more than the other four; even after the effect was corrected for regression artifacts (Campbell and Ehrlebacher, 1970). The mean growth for the lowest I.Q. block is 17 I.Q. points over the 4 year period; and 16 I.Q. points for the 70-90 I.Q. block.

These low I.Q. students are demonstrating one year's growth for each year in school in mathematics, and even higher levels in reading decoding. They also show growth in the general language and conceptual skills assessed by the Slosson Intelligence Test, representing a net gain of 17 I.Q. points. The picture is less promising in the area of reading comprehension and vocabulary - these children need more systematic instruction in vocabulary concepts and word meanings (Baxter, 1977). However, in general these secondary analyses document the effectiveness of a mainstreamed model with direct instruction teaching techniques and materials and mastery learning principles for children who would be

[1] This study was replicated with a somewhat larger sample of low I.Q. children from the South and Southwest (where there was no kindergarten). Thus, the children were exposed to only 3 years of Direct Instruction. In this case, there was a significant interaction in Math; students with entry I.Q.s below 70 gained significantly less than the other groups during the third grade (Figure 3). This would seem to argue heavily for the importance of preschool/kindergarten education, especially for those entering with low skills.

typically considered "at risk" or mildly handicapped.

DOCUMENTATION ACTIVITIES

Resnick (1979) distinguishes between evaluation research - where the effectiveness of an instructional program or instructional variable is assessed using a valid design strategy - and documentation activities where test scores describing the impact of a program are recorded without the benefit of the comparison standards present in experimental or quasi-experimental designs. This appears to be a valid distinction. In the last decade there has been a good deal of documentation of the impact of Direct Instruction programs with handicapped children, particularly in Australia. These studies are summarized in Becker, Engelmann, Carnine and Maggs (1981), Maggs and Maggs (1980) and Stevens (1980). This type of activity is valuable for several reasons. It offers administrators, decision makers, and parents some evidence of the impact of these programs on children. It also offers supervisors and curriculum developers some strong indication of areas of program strength and weakness. Furthermore, when documentation activities are coupled with discussions with teachers and supervisors on the relevance of particular tests for assessing the content of the curriculum, and the appropriateness and fairness of the testing formats, evaluation practices can be improved.

Finally, though documentation activities are not research, per se, they can lead to solid ideas for areas of serious empirical inquiry. Brief descriptions of the documentation activities follow.

The documentation studies have included a variety of handicapping

conditions and learning situations, as well as subject areas taught. Several Australian studies document direct instruction's effectiveness with the mildly handicapped. Ten severely learning disabled elementary students who were taught with Distar averaged 15 months gain in 10 weeks of instruction on various achievement tests. They also made significant gains on the WISC and ITPA (Maggs, 1978). Two schools with 130 disadvantaged Anglo and aboriginal elementary students were taught with Distar materials. The incidence of learning disabilities at the schools was reduced, and after three years of instruction most children achieved higher academic scores than the norms for their age (Hawkes, Maggs and Moore, Note 1). Similarly, the efficacy of a direct instruction corrective reading program was documented with 14 students who were considered reading failures (Maggs and Murdoch, 1979). The average gains were 10 months in comprehension and 12 months in decoding during seven months of instruction.

Distar use has been documented in mainstreamed settings, also. Third graders in a large classroom, who exhibited a six year range in academic skill level, were taught with direct instruction methods for a year. Groups of students with severe, mild, and no skill deficits all improved dramatically on achievement tests. There was no difference among the groups in rate of learning for the year (Maggs, 1978).

Generalization of skills learned from direct instruction has been documented by Maggs and Morath (1976). In a special school for the mildly retarded that had been using direct instruction for several years, direct instruction students were able to transfer learned skills to

settings outside the classroom more effectively than a control group.

In addition, Australian studies have documented the effect of direct instruction on more severely handicapped individuals. Thirty-three moderately retarded institutionalized youths (I.Q.s 35-55, ages 8-14) were taught with Distar Language and Reading for 32 months. These children averaged a 34 month gain on the Baldie Language Ability Test (Booth, 1978).

Recently, Apffel, Kelleher, Lilly and Richardson (1980) documented the effectiveness of two structured reading programs - Distar and Rebus (Woodstock, 1967) for 60 moderately retarded children. They wished to document their informal observations that given appropriate curricula material, many moderately retarded children can be taught to read on a functional level. Rebus and Distar were selected because both were structured programs, though Rebus utilized a whole word (rather than phonics_ approach and more independent seatwork, as opposed to oral group instruction. The authors did not use random assignment, and utilized criterion-referenced tests to assess progress. Thus, this cannot be considered an experimental design.

Yet the project did document that nearly all the young TMR participants demonstrated some capacity to profit from reading instruction. A few failed, but some students in both Distar and Rebus advanced to more conventional reading programs (p. 235). Using Lyndlsey's (1964) graphic technique for analysis of progress rate, the authors concluded that 20 of the 31 Distar students made Good Progress (65%) and the remaining 11 poor progress. For the Rebus sample, 6 of the 23 (26%) were considered to be making Good Progress and 11 of the 23 (48%) to be making poor progress. (The remaining were considered "neutral"). Despite some

confusing methodology, this study uses teacher-collected progress data to document that reading can be taught to many of the moderately retarded participants with a reasonably high success rate, and that direct instruction procedures are effective 65 percent of the time. The success rate for the whole-word approach is appreciably lower, as one would expect (Chall, 1979; Wolhberg, 1981). This is the sort of documentation that can be extremely useful for policy decisions.

NEW DIRECTIONS

The preceding studies fall into three major classes. The first group compared the effectiveness of various Direct Instruction programs with more traditional programs to assess their relative effectiveness. The second group, the norm-referenced comparisons, systematically evaluated the longitudinal growth rate of handicapped students involved in direct instruction programs with the academic growth rates of non-handicapped students. The final series of field studies documented the effects of various direct instruction programs with handicapped populations. Generally, a quite optimistic picture emerges. It would seem important to continue these types of activities for a variety of reasons, including the need to supply policy makers with hard, empirical data on effective programs.

But, in terms of research, I see a need to be more precise about the independent variables--to look carefully at isolated components of direct instruction (e.g., mastery learning, systematic use of corrections, the explicit teaching of problem solving strategies) and to determine

their relative contributions to student learning. This should be done using both the Distar and Corrective Reading Programs and other cirriculum materials. Direct instruction techniques can be used with other types of instructional programs, and it is essential to isolate which teaching techniques work best in adaptation.

I see two directions for these future programmatic research studies. The first is continuing formative evaluation studies such as those conducted by Carnine (in press), Falco, Gersten, Carnine and White (1981), and White (1979). In these studies, one variable is manipulated and the effects on student performance analyzed. Research to date has shown the superiority of teaching sequences that separate the presentation of visually or auditorily similar material (such as "b-d," "p-q") over those that attempt to present both members of the set at once (Carnine, in press; White, 1979). White (1979) demonstrated how training sequences for normal learners must be simplified and adapted for handicapped learners, who get confused with too much variation in irrelevant stimulus features when learning a new geometric concept. Falco, et al. (1981) demonstrated that, when teaching language concepts, dynamic presentations--using manipulations of real objects--are markedly superior to teaching from a series of line drawings. Other studies of this type are necessary to help isolate exact features of direct instruction programming that assist student learning.

A second, newer line of research, called instructional dimensions research (Cooley and Leinhardt, 1980), can also lend to important information about effective teaching practices. For this type of study,

classroom observers look at a set of variables of interest--degree of structure, allocation of academic time, student engagement during lessons, difficulty level of lessons—in a variety of classrooms. These variables are quantified, and the researchers attempt to explore relationships between them and student learning rate. Instructional dimensions studies, unlike the formative evaluation studies discussed above, allow research to effective combinations of variables in natural settings. They are an obvious outgrowth of the coarser, large-scale evaluation studies summarized in this paper and offer some hope of isolating effective special education teaching practices.

References

Apffel, J. A., Kelleher, J., Lilly, M. S., & Richardson, R. Developmental reading for moderately retarded children. Developmental Reading, 1980, 229-235.

Bateman, B. D. Teaching reading to learning disabled children: A fourth approach. In L. B. Resnick & P. A. Weaver (Eds.), Theory and practice of early reading (Vol 1). Hillsdale, NJ: Lawrence Erlbaum Associates, 1979.

Bateman, B., & Carnine, D. W. Direct instruction DISTAR. In N. G. Haring & B. Bateman, Teaching the learning disabled child. Englewood Cliffs, NJ: Prentice Hall, 1977.

Beard, R. M. An outline of Piaget's developmental theory for students and teachers. New York: Basic Books, 1969.

Becker, W. C., & Engelmann, S. Teaching 3: Evaluation of instruction. Chicago: Science Research Associates (SRA), 1976.

Becker, W. C., Engelmann, S., Carnine D., & Maggs, A. Direct instruction technology--Making learning happen. In P. Karoly & J. J. Steffen (Eds)., Advances in child behavior analysis and therapy (Vol 2.). New York: Gardner Press, 1980.

Bereiter, C., & Engelmann, S. Teaching disadvantaged children in the preschool. Englewood Cliffs, NJ: Prentice-Hall, 1966.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations in compensatory education tend to underestimate effects. In J. Hellmuth (Ed.), Disadvantaged child: Compensatory education--A national debate (Vol. 3) New York: Brunner/Mazel, 1970.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitfruit-multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

Carnine, D. Reducing problems associated with visually and auditorily similar correspondences. Learning Disabilities Quarterly, in press.

Carnine, D., & Gersten, R. (Eds.) Formative research studies on Direct Instruction. Eugene, OR: University of Oregon, 1979.

Chall, J. S. The great debate: Ten years later, with a modest proposal for reading stages. In L. B. Resnick & P. A. Weaver (Eds.), Theory and practice of early reading (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates, 1979.

Charters, W. W., & Jones, J. On the risk of appraising non-events in program evaluation. Educational Researcher, 1973, 2(11),

Cook, T. D., & Campbell, D. T. Quasi-experimentation: Design and analysis for field settings. Chicago: Rand McNally College Publishing Company, 1979.

Cooley, W. W., & Leinhardt, G. The instructional dimensions study. Education Evaluation and Policy Analysis, 1980, 2, 7-25.

Cooley, W. W. & Lohnes, P. Evaluation research in education. New York: Irvington, 1976.

Cronbach, L. J., et al. Toward reform of program evaluation. San Francisco: Jossey Bass, 1980.

Engelmann, S. Sequencing cognitive and academic tasks. In R. Kneedler & S. G. Tarver (Eds.), Changing perspectives in special education. Columbus, OH: Merrill, 1977.

Engelmann, S., & Carnine, D. Theory of instruction. New York: Irvington Press, in press.

Falco, R., Gersten, R., Carnine, D., & White, W. Enhancing attention of handicapped and non-handicapped students through a dynamic presentation of instructional stimuli. Unpublished manuscript. Eugene, OR: University of Oregon, 1981.

Fullan, M. The relationship between evaluation and implementation in curriculum. In A. Lewy (Ed.), Evaluation Roles, in press.

Gall, M. D. The importance of context variables in research on teaching skills. Research in Teacher Education, 1977, 28, 43-48.

Gersten, R., Becker, W. C., Heiry, T. J., & White, W. A. T. Relationship between entry IQ and yearly academic gains in a direct instruction model. Paper presented at AERA; Boston, Mass., April 1981.

Gersten, R., Carnine, D. W., & Williams, P. Measuring implementation of a structured educational model in an urban school district: An observational approach, in press.

Gersten, R., & Maggs, A. Five year longitudinal study of moderately retarded children in a direct instruction program. Unpublished manuscript. Eugene, OR: University of Oregon, 1981.

Gibson, E. J., & Levin, H. The psychology of reading. Cambridge, MA: Massachusetts Institute of Technology Press, 1975.

Gill, E. J. Methods of teaching reading: A comparison of results. In C. Walcutt, J. Lampart, T. McCracker (Eds.), Teaching reading: A phonic/linguistic approach to developmental reading. New York: MacMillan, 1974.

Gilmore, J. V., & Gilmore, E. C. Gilmore oral reading test. New York: Harcourt Brace Jovanovich, 1968.

Hallahan, D. P., & Kauffman, J. M. Introduction to learning disabilities: A psycho-behavioral approach. Englewood Cliffs, NJ: Prentice-Hall, 1976.

Horst, D. P., Tallmadge, G. K., & Wood, C. T. A practical guide to measuring project impact on student achievement. Monograph No. 1 on Evaluation in Education. Washington, D.C.: U. S. Government Printing Office, 1975.

House, E. R., Glass, G. V., McLean, L. D., & Walker, D. E. No simple answer: Critique of the "Follow Through" evaluation. Harvard Educational Review, 1978, 48, 128-160.

Kennedy, M. (Chair). Minutes of the proceedings of the Special Interest Group on Special Education Evaluation, American Education Research Association. Boston, Mass., 1980.

Kephart, N. C. The perceptual motor aspect of learning disabilities. Journal of Exceptional Children, 1964,

L'E. Stein, C., & Goldman, J. Beginning reading instruction for children with minimal brain dysfunction. Journal of Learning Disabilities, 1980, 13, 219-222.

Lindsley, O. P. Direct measurement and prosthesis of retarded behavior. Journal of Education, 1964, 147, 62-81.

Lloyd, J., Cullinan, D., Heins, E. D., Epstein, M. H. Direct instruction: Effects on oral and written language comprehension. Learning Disabilities Quarterly, 1980, 3, 70-76.

Lloyd, J., Epstein, M. H., & Cullinan, D. Direct teaching for learning disabilities. In J. Gottlieb & S. Strichart (Eds.), Perspectives on handicapping conditions: Current research and application in learning disabilities, in press.

Loucks, S., & Hall, G. E. A developmental model for determining whether the treatment is actually implemented. American Educational Research Journal, 1977, 14, 264-276.

Matarzzo, J. D. Wechler's measurement and appraisal of adult intelligence. New York: Oxford University Press, 1972.

Meyer, L. A. An experimental study of the treatment effects of word-analysis and word-supply correction procedures during word attack. Unpublished doctoral dissertation. Eugene, OR: University of Oregon, 1979.

Molitor, J., Watkin, N., Napior, D., & Proper, E. C. Education as experimentation--The non-Follow Through study. Cambridge, MA: Abt Associates, 1977.

Ozenne, D. et al. United States Office of Education. Annual evaluation report on programs administered by the U.S. Office of Education, FY 1975. Washington, D.C.: Capital Publ., Educational Resources Division, 1976.

Pflaum, S. W., Wahlberg, H. J., Karregianes, M. L., & Rahser, S. P. Reading instruction: A quantitative analysis. Educational Researcher, 1980, 9. 12-18.

Resnick, L. Personal communication. 1979

Rosenshine, B. V., & Berliner, D. C. Academic engaged time. British Journal of Teacher Education, 1978, 4, 3-16.

Serwer, B. L., Shapiro, B. J., & Shapiro, P. S. The comparative effectiveness of four methods of instruction on the achievement of children with specific learning disabilities. Journal of Special Education, 1973, 7, 241-249.

Slosson, R. L. Slosson intelligence test. East Aurora, NY: Slosson Educational Publications, 1963.

Stephens, E. Direct instruction research with moderately and severely handicapped in Australia: A review. ASET Journal, 1980, 12, 41-46.

Stevens, R., & Rosenshine, B. Advances in research on teaching. Exceptional Education Quarterly, in press.

Tallmadge, G. K. JDRP ideabook. Washington, D. C.: U.S. Government Printing Office, 1977.

Woodcock, R. W. Peabody rebus reading program. Circle Pines, MN: American Guidance Service, 1967.

Zoref, L. S. A prototype of how to evaluate implementation of a structured educational program. Unpublished doctoral dissertation. Eugene, OR: University of Oregon, 1981.